Enhanced Diffuse Field Model for Ad Hoc Microphone Array Calibration

Mohammad J. Taghizadeh[a,b], Philip N. Garner[a], Hervé Bourlard[a,b]

Emails: {mohammad.taghizadeh, phil.garner, herve.bourlard}@idiap.ch

[a] *Idiap Research Institute, Martigny, Switzerland*
[b] *École Polytechnique Fédérale de Lausanne (EPFL), Switzerland*

Abstract

In this paper, we investigate the diffuse field coherence model for microphone array pairwise distance estimation. We study the fundamental constraints and assumptions underlying this approach and propose evaluation methodologies to measure the adequacy of diffuseness for microphone array calibration. In addition, an enhanced scheme based on coherence averaging and histogramming, is presented to improve the robustness and performance of the pairwise distance estimation approach. The proposed theories and algorithms are evaluated on simulated and real data recordings for calibration of microphone array geometry in an ad hoc set-up.

Keywords: Ad hoc microphone array calibration, Diffuse field coherence model, Adequacy of diffuseness

1. Introduction

Microphone arrays are widely used in meeting rooms and teleconferencing applications. They are specifically employed to improve the speech quality by steering the beampattern towards a desired speaker [1, 2]. A plethora of applications includes distant speech recognition [3, 4], speaker localization [5, 6] and speech separation [7]. Recent advances in mobile computing and communication technologies enable use of cell phones, PDAs or tablets as an ad hoc microphone array. However, at the core of steered high quality acquisition, traditional localization and beamforming techniques are impractical without sufficient prior information on the microphone array geometry. Hence, in order to enable the effective use of ad hoc microphones for sound applications, calibration of the microphone array is required.

State of the art calibration techniques can be grouped into three categories. The first approach relies on transmitting a known signal to perform microphone calibration. Sachar et al. [8] presented an experimental setup using a pulsed acoustic excitation generated by five domed tweeters. The transmit times between speakers and microphones were used to find the relative geometry. Raykar et al. [9] used a maximum length sequence or chirp signal in a distributed computing platform. The time difference of arrival of the microphone signals were then computed by cross-correlation and used for estimating the microphone locations. Since the original signal is known, these techniques are robust to noise and reverberation.

The second category enables using an unknown signal; the microphone calibration is usually integrated with source localization. Flanagan and Bell [10] proposed a method using the Weiss-Friedlander technique, where the sensor location and direction of arrival of the sources are estimated alternately until the algorithm converges. Another approach was proposed by Chen et al. [11] where they introduced an energy-based method for joint microphone calibration and speaker localization. The energy of the signal is computed and a nonlinear optimization problem is formulated to perform maximum likelihood estimation of the source-sensor positions. This method requires several active sources for accurate localization and calibration.

McCowan et al. [12] proposed a calibration method based on the characteristics of a diffuse sound field model. A diffuse field can be roughly described as an acoustic field where the signals propagate with equal probability in all directions with the same power. The diffuse field is verified for meeting rooms and car environments [13] and it enables application of well-defined mathematical models for analysis of the acoustic field recordings. A particular property related to diffuse field recordings is the coherence function between pairwise microphone signals which is defined by a sinc function of the distance between the two microphones. Thereby, we can estimate the pairwise distances by least-squares fitting the computed coherence with the sinc function. This procedure is accomplished for each frame independently. To increase the robustness, the frame-based estimates are combined using k-means clustering [12]. This approach is applicable in a general room without the need for any explicit initialization or activating calibration signals. The study presented in this paper is built on the idea of incorporating the properties

of a diffuse field for ad hoc microphone array calibration.

The diffuse field has been studied rather extensively by many researchers with the aim of developing practical strategies for determining sound power, absorption measurements, and transmission loss. However, very few studies consider applicability of the associated models for microphone calibration. The purpose of this paper is to investigate the fundamental hypotheses of the diffuse field model and to elucidate the limitations and the scope of its applicability. The study of sound fields in lightly damped enclosed spaces can be approached in two different ways. One is based on solving the wave equation with known boundary conditions, which leads to descriptions in terms of the modes of the room. The other approach relies on statistical models for analysis of the field and requires far less information about the room geometry. We apply both of these methods to highlight the requirements for application of the diffuse field model to enable microphone array calibration.

The paper is organized as follows: The fundamentals of diffuse fields are studied in Section 2. We overview the characteristics and models of the diffuse field and the measurement for diffuseness. The methods to enhance the diffuse sound model are proposed in Section 3 and applied in the framework of microphone array calibration in Section 4. The fundamental limitation of the diffuse model are explained in Section 5. The experimental analyses are presented in Section 6 and the conclusions are drawn in Section 7.

2. Diffuse Field Fundamentals

2.1. Definition of Diffuse Field

A diffuse field is defined as an acoustic field consisting of a superposition of an infinite number of sound waves traveling with random phases and amplitudes such that the energy density is equivalent at all points. More precisely, all points in the field radiate equal power and random phase sound waves, with the same probability for all directions, and the field is homogeneous and isotropic [14]. A diffuse field can be realized if a point source is active in a highly echoic room. By removing the direct sound and the initial reflections from a recording of the sound, the remaining part consists of diffuse reflections. In addition, ambient distributed sound sources

yield a diffuse field, while the interference phenomena near the room boundaries and corners raise the energy level and reduce the diffuseness. In a free space, having many uncorrelated sources distributed at long distances can generate a diffuse field.

The diffuse sound field at its theoretical level does not exist in practice. However, in many cases, a diffuse sound field can be a useful approximation of the real acoustic field in an enclosure. The important point is then to measure the amount of diffuseness and evaluate its adequacy for different applications. The analytic studies consider two points of view: (1) the wave equation based approach that describes diffuse field through the modes in a room and (2) the statistical approach by considering an infinite number of free propagation plane waves, referred to as the plane wave model.

2.2. Diffuse Field Model

2.2.1. Mode Model

This theory analyzes a room as a pack of resonators with bandwidth proportional to the absorption of the walls [15, 16]. The 3 dB bandwidth of the mode is given by

$$B_{3\mathrm{dB}} = \frac{1}{2\pi\tau}, \tag{1}$$

where τ corresponds to the decaying time constant of the sound field energy [17].

By solving the equations of a homogeneous sound field with boundary conditions, we extract normal modes for the room. Each mode indicates a resonance frequency, and the distribution of these frequencies is determined by the shape and dimension of the room [18, 19]. At high frequencies f, the mode density depends solely on the room volume V as expressed through

$$\gamma(f) = \frac{4\pi V}{c^3} f^2, \tag{2}$$

where c denotes the speed of sound. The modal overlap is defined as the average number of

modes excited by a pure tone, and it is given by

$$\eta(f) = \gamma(f)B_{3dB} = \frac{4\pi V}{c^3} f^2 \frac{1}{2\pi\tau} \tag{3}$$

If the pure tone is close to the frequency of the mode, within a bandwidth of $2.2/T_{60}$, the adjacent mode is excited; T_{60} is equal to the time required for the level of a steady sound to decay by 60 dB after the sound has stopped. If $\eta(f) \geq 3$[1], there are enough excited modes to generate a diffuse field in the room [20], hence the critical frequency to achieve diffuseness is obtained as

$$f_s = \sqrt{\frac{3c^3\tau}{2V}} \tag{4}$$

This frequency is known as the Schroeder frequency [21, 22].

2.2.2. Plane Wave Model

An alternative analysis approach, which does not need acoustic information, relies on a statistical model. In the plane wave model or the statistical model, a diffuse field is defined by the superposition of a large set of plane waves impinging from all directions. We consider the steady state sound field generated by a pure tone source in a reverberant room. The time domain sound pressure $P(t)$ at a point far from the walls and the source is expressed as

$$P(t) = \lim_{q\to\infty} q^{-1/2} \sum_{i=1}^{q} b_i \cos(\omega t + \varphi_i), \tag{5}$$

where b_i and φ_i are random variables and independent of each other; φ_i has a uniform distribution in $[0, 2\pi]$ and b_i has a normal distribution; ω denotes the angular frequency and q is the number of plane waves. Each point in the field receives sound pressure from all directions [21]. Considering

[1] Deriving the 3D modes in a rectangular room, a decomposition of an oblique mode into eight plane waves can be obtained. Hence, for Υ model overlap, we get 8Υ plane waves. Some heuristics indicate that 24 plane waves is a lower bound for generating diffuse sound, therefore $\Upsilon = 3$ is the smallest value to achieve diffuseness as considered in Schroeder frequency (4).

this spatial uniformity, we can compute an average sound pressure through

$$P(t) = \lim_{q,m\to\infty} (qm)^{-1/2} \sum_{j=1}^{m} \sum_{i=1}^{q} b_{ij} \cos(\omega t + \varphi_{ij}), \tag{6}$$

where m is the number of different directions from which plane waves impinge on a point in the field. In three dimensions, the distribution of the plane waves is such that there is at least one plane wave at each $4\pi/m$ steradian. The plane wave model is particularly useful at medium to high frequencies; it requires no details about the room geometry. The accuracy however, degrades at low frequencies and the effects of interference is ignored. Waterhouse [23] extended this approach by considering the interference phenomena that occur near the walls. The studies in this paper rely on the basic mode model and the plane wave model.

3. Enhanced Diffuse Field Model

3.1. Averaging the Coherence Function

3.1.1. Cross Correlation

The correlation function of the sound pressures at two points in an acoustic field is defined as

$$C = \frac{\int_0^T P_1(t)P_2(t)dt}{\sqrt{\int_0^T P_1^2(t)dt \int_0^T P_2^2(t)dt}}. \tag{7}$$

The cross correlation function in a diffuse field has a closed form analytic solution [24, 25]. Suppose a plane wave passes two points located on the z-axis with separation d, the correlation function would be $\cos(\kappa d \cos\phi)$ where κ is the wavenumber and ϕ is the polar angle defined as the angle between the wave front and the line connecting the two points [26]. The value of C for a diffuse field can be obtained by averaging the cross correlation for all directions, as

$$\begin{aligned} C &= \int_0^{\pi} \int_0^{2\pi} \cos(\kappa d \cos\phi) \sin\phi \, d\theta \, d\phi / 4\pi \\ &= \sin(\kappa d)/(\kappa d), \end{aligned} \tag{8}$$

where θ is the azimuth angle.

3.1.2. Coherence Averaging

We consider a scenario in which n microphones record a diffuse field pressure signal. Suppose that S_i and S_l represent the spectral representation of the signals in Fourier domain at microphones i and l respectively. The cross spectral density is

$$\Phi_{il}(\omega) = S_i(\omega)S_l^*(\omega), \tag{9}$$

where "$*$" is the conjugate transpose operator. The coherence of two signals is the cross spectrum normalized by the square roots of the auto spectra, defined concisely as

$$\Gamma_{il}(\omega) = \frac{\Phi_{il}(\omega)}{\sqrt{\Phi_{ii}(\omega)\Phi_{ll}(\omega)}}\,. \tag{10}$$

In a perfect diffuse field, at each frequency component, the coherence is a sinc function, which holds if long time averaging (7) is taken [27]. As the frequency analysis is conducted on short frames, we propose to collect several frames and take an average over the frame-based coherence to achieve an estimate conforming to the sinc model. Therefore, we define an average coherence function as

$$\tilde{\Gamma}_{il}(\omega) = \frac{1}{J}\sum_{j=1}^{J} \Re\left(\Gamma_{il}^{j}(\omega)\right) = \mathrm{sinc}\left(\frac{\omega d_{il}}{c}\right), \tag{11}$$

where the operator $\Re(.)$ takes the real part of its argument; d_{il} is the distance between the two microphones, j denotes the frame index and J is the total number of frames. Based on this model, estimation of the distance between two microphones is possible by fitting a sinc function to the coherence of their signals. The conventional approach applies sinc function fitting on a frame-basis [12]. The theory asserted in this section suggests that an averaging method can improve pairwise distance estimation. We elaborate on the empirical evidence to verify this hypothesis in Section 6.

3.2. Boosting the Power

The theory of diffuse field analysis is developed under the assumption that the contribution of air absorption to the total enclosure absorption is negligible. In a silent room, where a diffuse field

is generated by the ambient sources such as running devices, computers, etc., the amplitude of the source signal is very weak. Therefore the prohibitive cost of air absorption affects the energy distribution. This condition tends to violate the necessary assumption of negligible energy loss during a mean free propagation. Hence, we propose to provide additional sources in a particular set up to boost the sound field power.

The diffuse field is better realized for high frequencies, as more modes are excited leading to an increase in the number of plane waves (Table 1). However the air absorption also increases with frequency; the acoustic intensity[2] of a plane wave as a function of the propagation distance r is expressed as

$$I(r, \omega) = I_0(\omega)\mathrm{e}^{-r/\xi(\omega)}, \tag{12}$$

where $I(r, \omega)$ is the intensity r meters from the source, $I_0(\omega)$ is the original intensity of the source with frequency ω and $1/\xi(\omega)$ is the attenuation factor, which increases with frequency. Therefore, if the source has a very low power, the high frequencies can diminish and the low frequencies, which do not excite enough resonance modes, remain in the sound field. This phenomenon reduces the diffuseness. Hence, we speculate that increasing the energy of the sound field yields higher diffuseness, and enables more accurate distance estimation. This idea has been evaluated empirically through the experiments conducted in Section 6.3.

3.3. Diffuseness Evaluation

3.3.1. Broadband Power Pattern

We consider a well-designed symmetric and regular spherical array of n microphones. The spectral representation of the signals recorded by microphone array in Fourier domain is denoted by $\mathcal{S}(\omega) = [S_1(\omega), S_2(\omega), \ldots, S_n(\omega)]^T$. Suppose that the beamformer weights steered towards direction $a(\theta, \phi)$ is represented by $\mathcal{F}(\omega, a(\theta, \phi)) = [F_1(\omega, a(\theta, \phi)), F_2(\omega, a(\theta, \phi)), \ldots, F_n(\omega, a(\theta, \phi))]$, the response of the array by applying the beamformer would be

$$Y(\omega, a(\theta, \phi)) = \mathcal{F}(\omega, a(\theta, \phi))\mathcal{S}(\omega). \tag{13}$$

[2]Sound power per unit area.

Given Y, the directional power can be measured as $Y^2(\omega, a(\theta, \phi))$. The directional power can be used to evaluate the level of diffuseness. As stated in Section 2, the power in a diffuse field is isotropic, which indicates equal power accumulated from all directions.

In a broadband diffuse field, we can apply a filter to improve the model fitting by restricting the broadband processing to frequencies conforming to the theoretical diffuseness bounds. The enhanced model can then be evaluated in terms of the isotropic power distribution using a broadband beamformer. Given Y, the beamformer output for the spectrum of signal, the broadband beampattern is given by

$$B(a(\theta, \phi)) = \Lambda \sqrt{\int_{\Omega} Y^2(\omega, a(\theta, \phi)) d\omega}, \tag{14}$$

where $\Omega = [\omega_{\min}, \omega_{\max}]$ is the frequency band of the signal and Λ is a normalization factor given by

$$\Lambda^{-1} = \max_{\theta, \phi} \sqrt{\int_{\Omega} Y^2(\omega, a(\theta, \phi)) d\omega}\ . \tag{15}$$

We can see that the broadband pattern can be interpreted as a weighted average of the beamformer's output over the broadband spectrum [28]. Accordingly, the broadband power-pattern would be

$$\mathcal{P}(a(\theta, \phi)) = |B(a(\theta, \phi))|^2. \tag{16}$$

3.3.2. Diffuseness Evaluation Measure

The appropriate application-specific criterion is necessary to evaluate the adequacy of the diffuseness. In this section, we propose a novel approach for evaluating the diffuseness in the room to assess the diffuseness adequacy for estimating pairwise distances. For the particular application of microphone calibration, a *pointwise* diffuseness is important, which indicates that the angular distribution of the power at any point is equal in all directions.

To measure the signal power, we propose to use a superdirective beamformer by steering the beam toward several representative directions of the space. In real scenarios, the ambient sound source in the environment does not have the same power at all frequencies, so it is crucial to consider the broadband power-pattern as explained in Section 3.3.1. After normalization,

we have to compare the three-dimensional (3D) pattern to a sphere of radius one. To obtain the broadband power-pattern at a particular point A in space, the microphone array has to be centered at A. Hence, the diffuseness level is defined as

$$\mathcal{X}_A = \frac{3}{4\pi} \iiint_{\mathcal{P}_A(\theta,\phi)} \rho^2 \sin\phi \; d\rho \; d\phi \; d\theta, \tag{17}$$

where $\mathcal{P}_A(\theta, \phi)$ is the measure of the power stated in (16) that is received from a direction with azimuth θ and polar angle ϕ in the diffuse field at point A; ρ denotes the radial distance in the Spherical coordinate system. $\mathcal{X}_A$ equals 1 if we have a complete diffuse field at point A.

Computation of $\mathcal{P}_A(\theta, \phi)$ is not easy and we need a 3D microphone array with a carefully designed symmetric and regular geometry. To simplify this computation, we consider reducing the 3D pattern to 2D by averaging over all angles ϕ. By defining $Q_A(\theta)$ as a 2D approximation of the 3D pattern $\mathcal{P}_A(\theta, \phi)$ through

$$Q_A(\theta) = \int_0^{\pi} \int_0^{\mathcal{P}_A(\theta,\phi)} \rho \; d\rho \; d\phi, \tag{18}$$

and

$$\nu = \max_{\theta} Q_A(\theta), \tag{19}$$

we derive $\widetilde{\mathcal{X}}_A$ as an approximation of $\mathcal{X}_A$ through

$$\widetilde{\mathcal{X}}_A = \frac{1}{\pi \nu^2} \int_0^{2\pi} \int_0^{Q_A(\theta)} r \; dr \; d\theta. \tag{20}$$

The approximated quantity $\widetilde{\mathcal{X}}_A$ is more practical, and it has enough accuracy for our application as we investigate numerically in Sections 6.3 and 6.4. The conventional methods consider mere sphericity and roundness to measure the level of diffuseness [29] whereas the proposed method is capable of directly measuring the isotropic sound field power at any given point in space; hence, the proposed diffuseness measure yields more accurate results.

4. Ad Hoc Microphone Array Calibration

4.1. Conventional Method

The following objective measure has been used to fit a sinc function for a broadband spectrum of coherence function and estimate the pairwise distance [12]

$$\delta_{il}^{j}(d) = \int_{\omega_{min}}^{\omega_{max}} \left| \Re\{\Gamma_{il}^{j}(\omega)\} - \mathrm{sinc}\left(\frac{\omega d}{c}\right) \right|^2 d\omega, \tag{21}$$

By minimizing $\delta_{il}^{j}(d)$ over d, we obtain an estimate $\tilde{d}_{il}^{j}$ per frame

$$\tilde{d}_{il}^{j} = \arg\min_{d} \ \delta_{il}^{j}(d). \tag{22}$$

The pairwise distance has been estimated for each frame of the sound signal. To improve the estimation accuracy, the estimates of multiple frames are combined using k-means clustering to remove the large-error estimates by grouping the points in two clusters. The clustering step is costly and requires long recorded signals to enable accurate estimation.

4.2. Proposed Averaging Method

The theoretical analysis carried out in Section 3.1 showed that the coherence function of a long segment is a sinc function and this model is not exact for a single frame. To obtain a sinc function model, we need to average over a sequence of frames. Figure 1, shows empirical evidence for this argument, and supports the requirement for averaging prior to fitting. The nonlinear characteristic and quick damping of the sinc function can lead to huge errors by only slight deviation from a diffuse field.

The averaging of the coherence of multiple frames prior to fitting the sinc function requires fewer frames than the clustering approach, and is very effective to improve the pairwise distance estimation performance. To state it more precisely, we consider J frames to extract the distance

between two microphones i and l. The averaging method is expressed as

$$\delta_{il}(d) = \int_{\omega_{min}}^{\omega_{max}} \left| \left[\frac{1}{J} \sum_{j=1}^{J} \Re\left(\Gamma_{il}^{j}(\omega)\right) \right] - \operatorname{sinc}\left(\frac{\omega d}{c}\right) \right|^2 d\omega \,,$$
$$\tilde{d}_{il} = \arg\min_{d} \ \delta_{il}(d) \tag{23}$$

4.3. Outlier Detection Techniques

In practice, there is no complete diffuseness and the characteristic of the sound field changes due to irregularities and acoustic ambiguities. This phenomenon results in outlier observations in the coherence function which lead to a high error in pairwise distance estimation. Hence, we propose to apply an outlier detection technique after averaging the coherence of multiple frames. The goal of outlier detection is to increase the quality and robustness of a data analysis approach.

We consider statistical outlier detection techniques based on k-means (parametric-based) as well as histogram (non-parametric) methods. In the parametric approach, we consider a profile and unsupervised learning with certain criteria to identify the outliers in pairwise distance estimation. More experiments show that the erroneous estimates do not conform to a specific parametric model. Hence, we resort to a non-parametric histogram-based approach. In the histogramming method, the outliers are identified through a fixed threshold. In addition, this method requires less memory and computational cost, although finding the optimal size of the bins for a large number of attributes is a challenging task. The experimental analyses conducted in sections 6.2 and 6.4 confirm the validity of the averaging method followed by outlier removal using histogram-based clustering for robust estimation of the pairwise distance. Furthermore, we show that histogram clustering outperforms the k-means clustering approach. At the final step in calibration of the microphones, the geometry is extracted using the s-stress method [30].

5. Fundamental Limitation of Diffuse Model

This section explains the fundamental limitations and the performance bound of distance estimation using a diffuse field coherence model. As we have already seen earlier in the paper, the

spatial coherence of two signals in a diffuse field is a sinc function of the pairwise distance (11). This function decreases quickly and, as shown in Figure 1, it disappears after one cycle. Hence, the coherence measured in the first cycle is vital in estimation accuracy.

We consider three scenarios, being a medium size room ($8 \times 5.5 \times 3.5\,\mathrm{m}^3$), a large size room ($24 \times 16.5 \times 10.5\,\mathrm{m}^3$) and a very large size room ($48 \times 33 \times 21\,\mathrm{m}^3$). The second zero crossings on the sinc function as expressed in (11) occur at 343 Hz, 114 Hz and 57 Hz for pairwise distances of 1 m, 3 m and 6 m, respectively. Hence, diffuseness at frequencies lower than these frequencies are important.

On the other hand, the Schroeder frequency is obtained as $f_s = \sqrt{6c^2/\alpha Z}$ where α is the average absorption coefficient of the walls with a surface area of Z [22]; therefore, for an average absorption coefficient $\alpha = 0.07$ and $c = 343$ m/s, the Schroeder frequencies for these three rooms are 235 Hz, 78 Hz and 39 Hz respectively. As indicated in Section 2.2.1, a diffuse field cannot be generated in a room with a monochromatic source under the Schroeder frequency.

The mode model can be used for computing the acoustic pressure in modal behavior. Diffuseness at each frequency band can be illustrated by expansion modes. Table 1 summarizes the number of modes for each one-third-octave band in three room sizes. Based on theory, we hypothesize that increasing the dimension of the room increases the diffuseness, in particular at low frequencies which are highly effective in distance estimation. In addition, by increasing the pairwise distances, the number of discrete frequencies below the second zero crossing decreases linearly so we speculate that a linear regression can illustrate the relationship between the errors and distances. The empirical evaluations carried out in Sections 6.3–6.6 confirm the validity of these hypotheses. These experiments enable formulating a relation between room dimension and achievable distance estimation.

6. Experimental Analysis

This section presents the numerical results to evaluate the proposed theories and hypotheses. The microphone calibration performance measure must be robust to rigid transformations (translation, rotation and reflection). Hence, we use the distance between the actual locations $\boldsymbol{X}$ and

estimated locations $\hat{\boldsymbol{X}}$ as defined in [31]

$$\begin{aligned} \text{dist}(\boldsymbol{X}, \hat{\boldsymbol{X}}) &= \frac{1}{n} \left\| \boldsymbol{L}\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{L} - \boldsymbol{L}\hat{\boldsymbol{X}}\hat{\boldsymbol{X}}^T\boldsymbol{L} \right\|_{\text{F}} , \\ \boldsymbol{L} &= \boldsymbol{I}_n - (1/n)\mathbf{1}_n\mathbf{1}_n^T, \end{aligned} \tag{24}$$

where $\|\cdot\|_{\text{F}}$ denotes the matrix Frobenius norm. The $\mathbf{1}_n \in \mathbb{R}^n$ is the all ones vector, $\boldsymbol{I}_n$ is the $n \times n$ identity matrix and $\boldsymbol{X}, \hat{\boldsymbol{X}} \in \mathbb{R}^{n\times\eta}$, where η is the dimension of the space. The distance measure stated in (24) is useful to compare the performance of different methods when the microphone array geometry is fixed.

6.1. Data Recording Set-up

6.1.1. Simulation Scenarios

We simulate a medium size room of dimensions $8 \times 5.5 \times 3.5\ \text{m}^3$, which has the same dimension of the room in the real scenario. The room is equipped with 48 omni-directional loudspeakers playing independent white Gaussian noise. These are divided into 3 uniform circular arrays with diameters of 1.5 m, 2.5 m and 1.5 m, producing the sound field. The three circular loudspeaker arrays are parallel to the floor and located at the center of the planar area of the room at 0.1 m, 1.75 m and 3.4 m height. A uniform 8-channel circular microphone array located at center of the room is used to record the sound field. The diameter of the array is adjusted such that the pairwise distance between the microphones is equal to $\{0.1, 0.2, 0.3, \ldots, 0.8\}$ m; that corresponds to the microphone array diameters of $\{0.26, 0.52, 0.78, 1.04, 1.30, 1.57, 1.83, 2.10\}$ m. To enable evaluations for larger distances beyond 0.8 m, the 8-channel array is replaced with a 16-channel uniform circular microphone array with a diameter equal to $\{0.9, 1, \ldots, 2\}$ m. Figure 2 depicts a top view of the simulated scenario.

In addition, for investigation of the effect of room dimension on diffuseness of the field, and distance estimation, a large room as well as a very large room of dimensions $24 \times 16.5 \times 10.5\ \text{m}^3$ and $48 \times 33 \times 21\ \text{m}^3$ such that each dimension is 3 and 6 times bigger than real scenario are simulated. The same set-up of loudspeakers are used where the diameters are expanded by a factor of 3 and 6 . The same microphone array is used to record the sound field. The diameter

of the array is adjusted such that the pairwise distance between the microphones are varied from 0.1m to 10 m.

The room impulse responses are generated with the image source model [32] using intra-sample interpolation up to 15^{th} order reflections. The corresponding reflection ratio, β used by the image model was calculated via Eyring's formula:

$$\beta = \exp(-13.82/[c \times (L_x^{-1} + L_y^{-1} + L_z^{-1}) \times T_{60}]), \tag{25}$$

where L_x, L_y and L_z are the room dimensions. The temperature of the room is assumed to be 20° Celsius, thus $c = 343$ m/s. In our experiments, $T_{60} = 300$ ms for the medium size room. The direct-path propagation is discarded from the impulse response for generating a diffuse sound field [5].

6.1.2. Real Data Scenario

In addition to the simulated recordings, we use the geometrical setup of the MONC corpus to record the sound field in a meeting room [33]. The enclosure is a 8×5.5×3.5 m^3 rectangular room and it is moderately reverberant. It contains a centrally located 4.8×1.2 m^2 rectangular table. Twelve microphones are located on a planar area parallel to the floor at height of 1.15 m: Eight of them are located on a circle with diameter 20cm and one microphone is at the origin. There are three additional microphones at a 70cm distance from the central microphone. The microphones are Sennheiser MKE-2-5-C omnidirectional miniature lapel microphones. The floor of the room is covered with carpet and surrounded with plaster walls and two big windows.

The recordings were made in two scenarios: (1) Collecting the diffuse sound field of ambient noise in the room without any additional source and (2) playing extra sounds by putting two small loudspeaker under the table, and covering them with anti-acoustic material, so that the direct paths between loudspeaker and microphones are prohibited to ensure diffuseness. The microphone placement is depicted in Figure 3. The sampling rate is 48 kHz while the processing applied for microphone calibration is based on a down-sampled signal at rate 16 kHz to reduce the computational cost. The experiments are conducted using $c = 343$ m/s that corresponds to

20° Celsius temperature of the room.

6.2. Averaged Coherence Function

Figure 1 shows a real data example of the coherence of one frame (top) and the coherence function averaged over 100 frames (bottom) along with the fitted sinc function. As we can see, averaging is crucial prior to fitting the model by least square regression. The conventional method [12] fitted a sinc function on a single frame followed by k-means clustering of multiple frames to determine the distance. The numerical results show that the error of fitting a sinc function on the averaged coherence function is 35 times smaller than the conventional method for small distances. Furthermore, this method speeds up the calibration by a factor of 60 compared to the k-means clustering method in terms of CPU time using the same number of frames.

6.3. Diffuseness Evaluation

The first experiments consider measuring the diffuseness with the method proposed in section 3.3.2. A superdirective beamformer was used for measuring the power of the received signal from all directions. Figure 4 shows the patterns for the simulated very large room at distances 2 m (top) and 5 m (bottom) from the room center. We can see that a more isotropic power is obtained if the point of measurement is closer to the room center. Based on the definition stated in (20), the diffuseness levels at 2m and 5m distances from the center of the room are .92 and .84 respectively, which shows that the diffuseness reduces as we get closer to the borders.

The diffuseness for the real data recorded at the meeting room without additional sources is measured as 0.70. We increased the power by playing white Gaussian noise from the two small loudspeakers put under the table. Figure 5 shows the pattern with the proposed sound field augmenting method compared to the initial recordings. A more isotropic sound field is obtained as the pattern is closer to a circle. Quantitatively, the diffuseness is improved to 0.83, that indicates a 19% increase in diffuseness level.

Based on the diffuseness level quantified in this section and the real data distance estimation results listed in Table 2 (explained further in the next Section 6.4), we can see that a diffuseness

level around 0.7 is a reasonably adequate diffuseness as we can estimate the pairwise distances with less than 5% relative error.

As discussed in Section 3.3.2, estimation of the directional power can be accomplished by a symmetric and uniform microphone array; that implies a carefully designed spherical (3D) or circular (2D) array. The 2D approximation reduces and simplifies some of the computations. We consider this level of approximation reasonable as the obtained calibration error and distance measurement are not very sensitive to the quantified diffuseness [14]. Furthermore, the numerical results confirm that the quantified diffuseness levels are in agreement with the distance estimation results (Table 2).

6.4. Distance Estimation Performance

In order to estimate the pairwise distances, two microphone signals are processed using a short time Fourier transform of 64ms frames obtained by applying the Tukey window with parameter = 0.25. The total length of each microphone signal is $30s$. For each frame, we compute the coherence function through (10) and estimate the pairwise distance by fitting a sinc function as stated in (21) and (22). In the baseline approach, each frame is processed independently, which yields 468 point estimates of pairwise distances. To obtain a single estimate of the distance between the two microphones, clustering is applied on the point estimates. Based on k-means clustering, the center of the cluster with the smaller error determines the pairwise distance [12]. Using our enhanced model elaborated in Section 3.1, a sinc function is fitted to the averaged coherence function.

We conduct the evaluations using simulated data in a controlled (almost ideal) diffuse field in the medium and large size rooms as described in Section 6.1. Figure 6 illustrates the results. We can see in Figure 6 (top: averaging method) that in the medium size room, the pairwise distances smaller than 1 m can be estimated with less than or equal to 0.02 m error (the 90% confidence interval is 0.03 m). The estimates become highly erroneous beyond 1 m. By using the conventional method (Figure 6 top: k-means clustering), observations show that this method is only applicable when the microphones are located in close proximity to each other (i.e., the

pairwise distance less than 30 cm). Figure 6 (bottom) illustrates that in the large size room, the averaging method is effective for estimation of pairwise distances up to 3 m.

The relative error for distance estimation d_i can be quantified as

$$\epsilon_i = \sqrt{\frac{\sum_{l=1}^{N} \left(\frac{\hat{d}_{il} - d_i}{d_i}\right)^2}{N}} \tag{26}$$

where $\hat{d}_{il}$ is l^{th} estimation of distance d_i and N is the number of microphone pairs with pairwise distance d_i. Figure 7 shows that measure of ϵ_i is almost constant for each room and we can fit a linear regression model on the relative error. As depicted in Figure 7 (top), the line corresponds to 0.0164 m relative error and the residual error of the linear regression is 0.0028 for the medium size room. In addition, we performed some evaluations in the large room set-up as described in Section 6.1. The theories of the sinc function coherence model hold for up to 3 m pairwise distance, which is also verified through our experiments in a diffuse field. Similar to the previous experiment, we can fit a linear regression model on the relative error as depicted in Figure 7 (bottom). The line corresponds to 0.0124 m relative error and the residual error of the linear regression is 0.0012. We can see that the following mathematical model holds for estimation of pairwise distance

$$\hat{d} \sim \mathcal{N}(d, (d\epsilon)^2) \tag{27}$$

where $\mathcal{N}$ denotes the normal distribution and ϵ is the mean of the relative errors in distance estimation which is equal to 0.0164 and 0.0124 in the medium and large size rooms respectively. A smaller ϵ indicates that diffuseness is better realized in the larger room.

We further conduct some evaluations using real data recorded in a meeting room. Figure 8 illustrates that, for microphones 7 and 8 located 7.6 cm apart, the k-means clustering estimated distance is 8.2 cm. Figure 9 (top) demonstrates that for microphones 11 and 5 where the distance is 77.38 cm, it is not possible to provide a reliable estimate by fitting a sinc function on a single frame and k-means clustering. The estimated distance is 66 cm which shows more than 11 cm (14%) error.

The proposed averaging method enables more accurate point estimates with fewer outliers. Figure 9 (bottom) shows fusion of the k-means method with an averaging technique for estimating the distance between microphones 11 and 5. The averaging is performed on each 5 frames with 80% overlapping. The results shows that the percentage of outliers is reduced so the estimated pairwise distance is 76.6 cm, amounting to 8 mm (1%) error.

Although the averaging method reduces the number of outliers, the k-means clustering is not stable and it can generate the wrong winner class. Figure 10 shows distance estimation for microphones 11 and 6. The estimated distance is 90.2 cm, whereas the correct distance is 80cm. The winner class is wrong using k-means clustering. We propose to remove the outliers using a histogram clustering method, which also offers computational speed advantages over the k-means algorithm. Furthermore, as discussed in Section 4.3, it is a more appropriate technique for removing outliers compared to k-means clustering. The two-dimensional histogram clustering is shown in Figure 11. Note that the histogram represents the difference of the positions (distance) of the microphones and not the positions themselves. This method is not dependent on the absolute position of the compared microphones. In the histogram method, the bin with the largest number of estimation points is the winner used for the final estimation. The resolution of the bins is a critical parameter for construction of the histogram. We observed empirically that a 50×50 histogram provides a good estimate; it corresponds to a resolution of an average 7mm in pairwise distance estimation. The two-dimensional histogram enables estimation of the pairwise distance as 80.3 cm which has only 3 mm error, equal to 0.4%. We can see that this method is more accurate than k-means clustering and it is more robust to noisy estimates in real data evaluations.

Table 2 summarizes all the results for pairwise distance estimation in the real data scenario. The first column is the ground truth distances. The second column is the root mean square error (RMSE) for the baseline method, and the third column is the RMSE for the boosted power diffuse field; it shows an improvement compared to the baseline. The fourth column corresponds to the results of using the averaging and two-dimensional histogram methods, which shows noticeable improvement. Applying this method on the boosted power diffuse field shows an additional slight improvement as listed in the last column. We can see that the averaging and two-dimensional

histogram are more important to achieve robust and accurate results. Furthermore, Figure 12 illustrates the measure of improvement using each method. We can see that although boosting the power increases the diffuseness, the improvement in pairwise distance estimation is small because measuring the diffuseness was done on all frequency bands, whereas only the low frequency part has contribution to the distance estimation. Therefore measuring the diffuseness at low frequencies is essential to predict the performance of distance estimation.

6.5. Array Calibration Performance

In the final section, we compare all methods for calibration of the geometry of the 9 microphones using real data. Figure 13 illustrates the microphone calibration results.

The geometry of the array is extracted using the state-of-the-art s-stress [31] method by solving the following optimization problem

$$\hat{\boldsymbol{X}} = \arg\min_{\boldsymbol{X}} \sum_{(i,j)\in E} \left(\left\| \boldsymbol{x}_i - \boldsymbol{x}_j \right\|^2 - \tilde{d}_{ij}^2 \right)^2 , \tag{28}$$

where $E \subseteq [n] \times [n]$ denotes the subset of the estimated pairwise distances and $\boldsymbol{x}_i$ represents the microphone location i. This method is a robust and accurate localization technique where the search space is constrained to the Euclidean geometry. The reconstruction error for the baseline method using the criterion stated in (24) is 8.83. The estimated error based on averaging method is 8.04.

To further improve the performance, we use the two-dimensional histogram to remove outliers. We can see the improved estimates using the hybrid of averaging method and outlier detection, where the averaging method is applied on five frames to estimate the pairwise distances and to construct the two-dimensional histogram; the estimated error is 5.00. Table 3 summarizes the results. The same number of frames is used by each method.

6.6. Diffuseness Adequacy for Pairwise Distance Estimation

The theory stated in section 2.2.1 asserts that the critical frequency to create a diffuse field is inversely proportional to the dimension of the room. Hence, as the room gets larger, the critical frequency gets smaller, and we can achieve a higher diffuseness especially at low frequencies. On the other hand, Equation (11) shows that by increasing the pairwise distance, the sinc function squeezes in the frequency domain; therefore, the diffuseness at low frequencies becomes highly important for fitting the coherence function and the estimated sinc function. Hence, estimation of large pairwise distances is difficult. Table 4 illustrates the relation between room dimension and maximum pairwise distance estimation.

As the simulation results illustrate, increasing the dimensions by a factor of 6 enables estimation of larger pairwise distances by a similar factor of 6. Therefore, in the room with dimensions $48 \times 33 \times 21\text{m}^3$, pairwise distances up to 6 m can be estimated accurately. Table 5 summarizes the results of distance estimation for the very large room.

Comparing the simulated and real data evaluations on the medium size room shows that, in a simulated as well as real diffuse field, we can estimate pairwise distance up to 1 m (Table 4).

Section 5 showed that between the second zero crossing frequency and the Schroeder frequency for the aforementioned three rooms (medium, large and very large), there are only two bands for distances 1m, 3m and 6m respectively (Table 1). Hence, the diffuseness generated by a tone is very weak and we may not be able to fit the sinc function to extract these pairwise distances.

In our particular case of using broadband signal, all bands that have more than 25 modes generate a diffuse field [20]. Our empirical evaluations show that at least 5 diffuse field bands below the second zero crossing frequency are necessary to achieve reasonable accuracy in distance estimation. Table 1 shows that in the medium size room, 5 bands (15–19) generate an adequate diffuse field at frequencies below the second zero crossing; similarly, in the large room and the very large room, the bands 10–14 and 7–11 generate adequate diffuse field distances corresponding to 1m, 3m and 6m respectively. Based on this theory and the second column of Table 1, estimation of 3 m distances in the medium size room are impossible. The experiments on

real data recordings confirm this theoretical insight. Hence, it becomes straightforward to determine the maximum distance which can be estimated using the diffuse field model. The procedure requires extraction of the modes for the room. The minimum frequency (f^*) to have 5 bands generating more than 25 modes lower than f^* yields the maximum resolvable distance as $d^* = c/f^*$. Hence, as f^* gets smaller (i.e. room gets larger) the maximum estimated distance is increased. The f^* is equal to 355, 112 and 56 Hz for the medium, large, and very large rooms respectively, cf. Table 1, 19, 14 and 11 band indices. Those correspond to the maximum distances of 0.97 m, 3.06 m and 6.12 m.

7. Conclusions

In this paper, we studied the diffuse field model to enable ad hoc microphone array calibration. The analyses showed the importance of averaging the coherence function prior to fitting the sinc function. The robustness was further improved using 2D histogram based clustering for outlier detection. The enhanced model was shown to outperform the conventional method significantly. The fundamental limitations of this approach were elaborated and effective strategies were proposed to enable estimation of array geometry in an arbitrary set-up. Based on the theoretical as well as empirical studies on adequacy of diffuseness, a mathematical relationship was characterized to link the room dimensions to the maximum resolvable distance using a diffuse field model. The theory explains why larger aperture arrays can be calibrated in larger enclosures and suggests a simple procedure to figure out the maximum distance that can be estimated using a diffuse field coherence model.

Acknowledgments

This work has received funding from the Swiss National Science Foundation under the National Center of Competence in Research (NCCR) on "Interactive Multi-modal Information Management" (IM2). The authors acknowledge the anonymous reviewers for the precise and helpful comments and remarks to improve the quality and clarity of the manuscript.

References

[1] Q. Zou, X. Zou, M. Zhang, Z. Lin, A robust speech detection algorithm in a microphone array teleconferencing system, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 5, 2001.

[2] R. Cutler, Y. Rui, A. Gupta, J. Cadiz, I. Tashev, L.-w. He, A. Colburn, Z. Zhang, Z. Liu, S. Silverberg, Distributed meetings: a meeting capture and broadcasting system, in: Proceedings of the tenth ACM international conference on Multimedia, 2002.

[3] M. L. Seltzer, Microphone array processing for robust speech recognition, in: PhD Thesis, Carnegie Mellon University, 2001.

[4] A. Asaei, H. Bourlard, P. N. Garner, Sparse component analysis for speech recognition in multi-speaker environment, in: The 11th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2010.

[5] M. J. Taghizadeh, P. N. Garner, H. Bourlard, H. R. Abutalebi, A. Asaei, An integrated framework for multi-channel multi-source localization and voice activity detection, in: IEEE workshop on Hands-free Speech Communication and Microphone Arrays, 2011.

[6] A. Asaei, M. J. Taghizadeh, M. Bahrololum, M. Ghanbari, Verified speaker localization utilizing voicing level in split-bands, Signal Processing 89(6), 2009.

[7] A. Asaei, M. J. Taghizadeh, H. Bourlard, V. Cevher, Multi-party speech recovery exploiting structured sparsity models, in: The 12th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2011.

[8] J. M. Sachar, H. F. Silverman, W. R. Patterson, Microphone position and gain calibration for a large-aperture microphone array, IEEE Transactions on Speech and Audio Processing 13(1), 2005.

[9] V. C. Raykar, I. V. Kozintsev, R. Lienhart, Position calibration of microphones and loudspeakers in distributed computing platforms, IEEE Transactions on Speech and Audio Processing 13(1), 2005.

[10] B. P. Flanagan, K. L. Bell, Array self-calibration with large sensor position errors, Signal Processing 81, 2001.

[11] M. Chen, Z. Liu, L. He, P. Chou, Z. Zhang, Energy-based position estimation of microphones and speakers for ad-hoc microphone arrays, in: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2007.

[12] I. McCowan, M. Lincoln, I. Himawan, Microphone array shape calibration in diffuse noise fields, IEEE Transactions on Audio,Speech and Language Processing 16(3), 2008.

[13] J. Bitzer, K. U. Simmer, K. Kammeyer, Theoretical noise reduction limits of the generalized sidelobe canceller (GSC) for speech enhancement, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999.

[14] T. Schultz, Diffusion in reverberation rooms, Journal of Sound and Vibration 16(1), 1971.

[15] W. Chu, Spatial cross-correlation of reverberant sound fields, Journal of Sound and Vibration 62(2), 1979.

[16] W. T. Chu, Eigenmode analysis of the interference patterns in reverberant sound fields, Journal of the Acoustical Society of America 68(1), 1980.

[17] T. Bravo, C. Maury, Enhancing low frequency sound transmission measurements using a synthesis method, Journal of the Acoustical Society of America 122(2), 2007.

[18] M. R. Schroeder, The schroeder frequency revisited, Journal of the Acoustical Society of America 99(5), 1997.

[19] M. Wankling, B. Fazenda, Studies in modal density-its effect at low frequency, in: Proceedings of the Institute of Acoustics, 2009.

[20] H. Nelisse, J. Nicolas, Characterization of a diffuse field in a reverberant room, Journal of the Acoustical Society of America 101(6), 1997.

[21] M. R. Schroeder, Measurement of sound diffusion in reverberation chambers, Journal of the Acoustical Society of America 31(11), 1959.

[22] M. R. Schroder, K. H. Kuttruff, On frequency response curves in rooms. comparison of experimental, theoretical and monte carlo results for the average frequency spacing between maxima, Journal of the Acoustical Society of America 76–80, vol. 34, 1962.

[23] R. V. Waterhouse, Interference patterns in reverberant sound fields, Journal of the Acoustical Society of America 27(2), 1955.

[24] C. F. Chien, W. W. Soroka, Spatial cross-correlation of acoustic pressures in steady and decaying reverberant sound fields, Journal of Sound and Vibration 48(2), 1976.

[25] B. Rafaely, Spatial-temporal correlation of a diffuse sound field, Journal of the Acoustical Society of America 107.

[26] C. Morrow, Point-to-point correlation of sound pressures in reverberation chambers, Journal of Sound and Vibration 16(1), 1971.

[27] R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, M. C. Thompson, Measurement of correlations coefficients in reverberan sound fields, Journal of the Acoustical Society of America 27, 1955.

[28] M. J. Taghizadeh, P. N. Garner, H. Bourlard, Microphone array beampattern characterization for hands-free speech applications, in: IEEE 7th Sensor Array and Multichannel Signal Processing Workshop, 2012.

[29] B. Gapinski, M. Grezelka, M. Ruck, The roundness deviation measurment with coordinate measuring machines, Engineering Review 26(2), 2006.

[30] T. F. Cox, M. A. A. Cox, Multidimensional scaling, Chapman-Hall, 2001.

[31] I. Borg, P. J. F. Groenen, Modern multidimensional scaling theory and applications, Springer, 2005.

[32] J. B. Allen, D. A. Berkley, Image method for efficiently simulating small-room acoustics, Journal of Acoustical Society of America 60(s1), 1979.

[33] The multichannel overlapping numbers corpus (MONC), Idiap resources available online:, `http://www.cslu.ogi.edu/corpora/monc.pdf`.

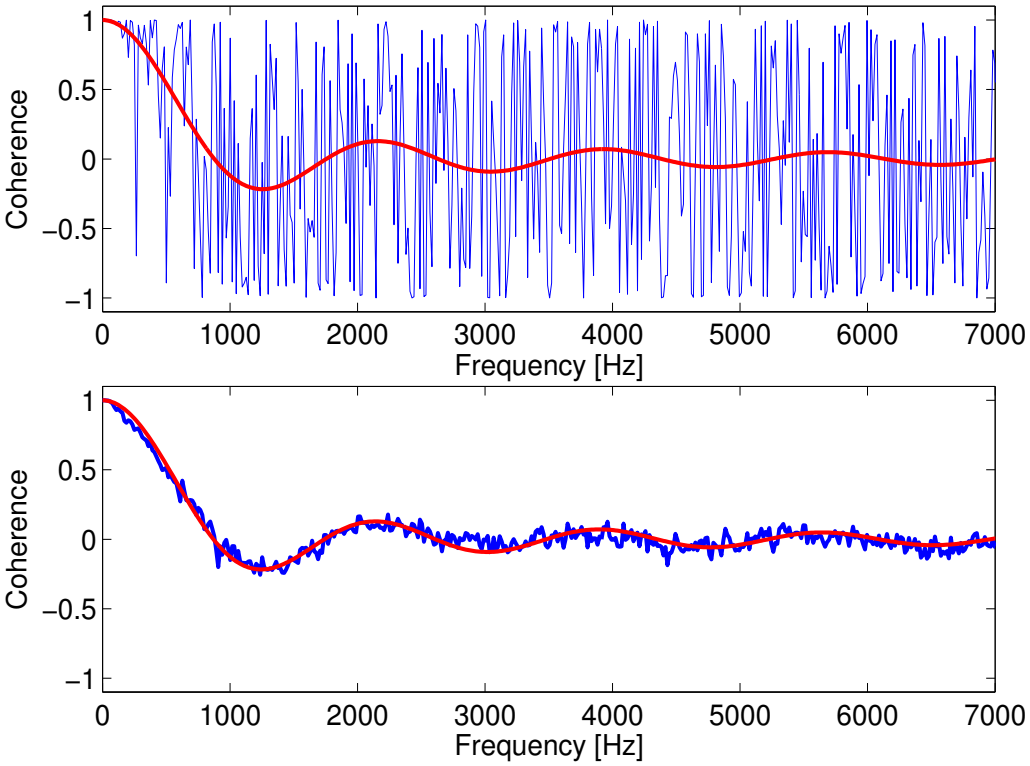

Figure 1: (Top) Fitting a sinc function (red) on one frame of diffuse field coherence (blue); the correct distance is 20 cm and the estimated distance is 19.3 cm. (bottom) Fitting a sinc function on average of 100 frames of diffuse sound field coherence; the estimated distance is 19.8 cm.

band index	$f_1 - f_2$ Hz	#modes/band (medium)	#modes/band (large)	#modes/band (very large)
1	4.44-5.6	0	0	1
2	5.6-7.1	0	1	2
3	7.1-9	0	0	3
4	9-11.2	0	1	6
5	11.2-14	0	1	6
6	14-18	0	4	20
7	18-22.4	1	6	25
8	22.4-28	0	6	53
9	28-35.5	1	19	100
10	35.5-45	2	29	205
11	45-56	3	50	340
12	56-71	7	100	734
13	71-90	9	205	1458
14	90-112	18	340	2589
15	112-140	30	684	5054
16	140-180	68	1508	11127
17	180-224	115	2589	15754
18	224-280	206	5054	39132
19	280-355	440	10611	82724

Table 1: Number of modes in the one-third-octave bands for medium size room ($8 \times 5.5 \times 3.5\,\mathrm{m}^3$), large size room ($24 \times 16.5 \times 10.5\,\mathrm{m}^3$) and very large size room ($48 \times 33 \times 21\,\mathrm{m}^3$).

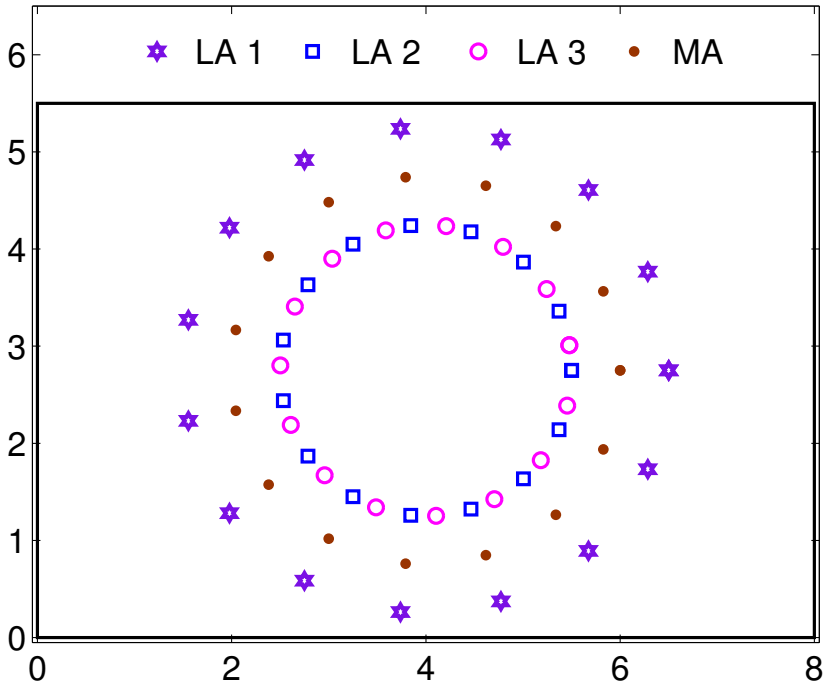

Figure 2: Top view of the simulated medium size room scenario: This scenario consists of three circular 16-element omni-directional loudspeaker arrays (LA) and one circular microphone array (MA) with the following set-up parameters: LA1 has diameter=2.5 m located at height=1.75 m; LA2 and LA3 have diameters=1.5 located at height=0.1 m and 3.4 m respectively. A 16-element microphone array is depicted with diameter=2 m and it is located at height=1.75 m. All arrays are parallel to the floor. The number of microphones and the diameter of the MA are varied as explained in Section 6.1.1 to generate various pairwise distances.

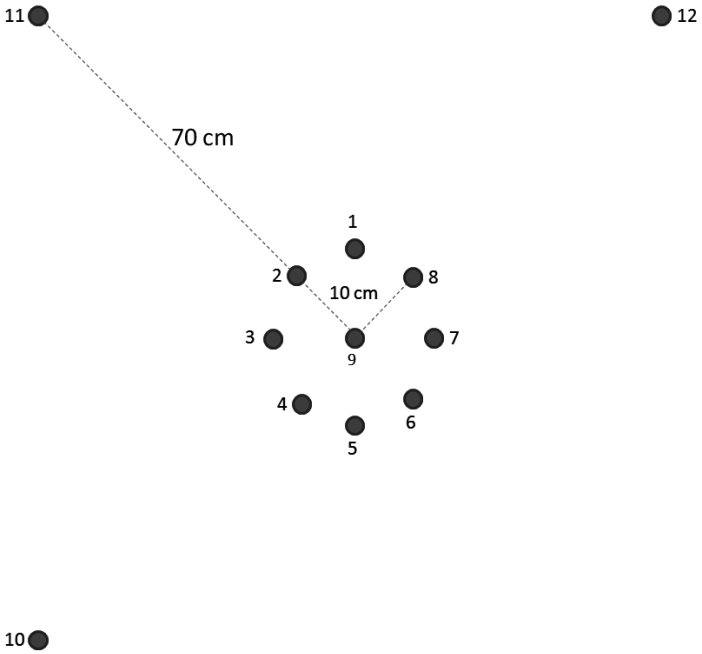

Figure 3: Microphone placement for real data recording scenario.

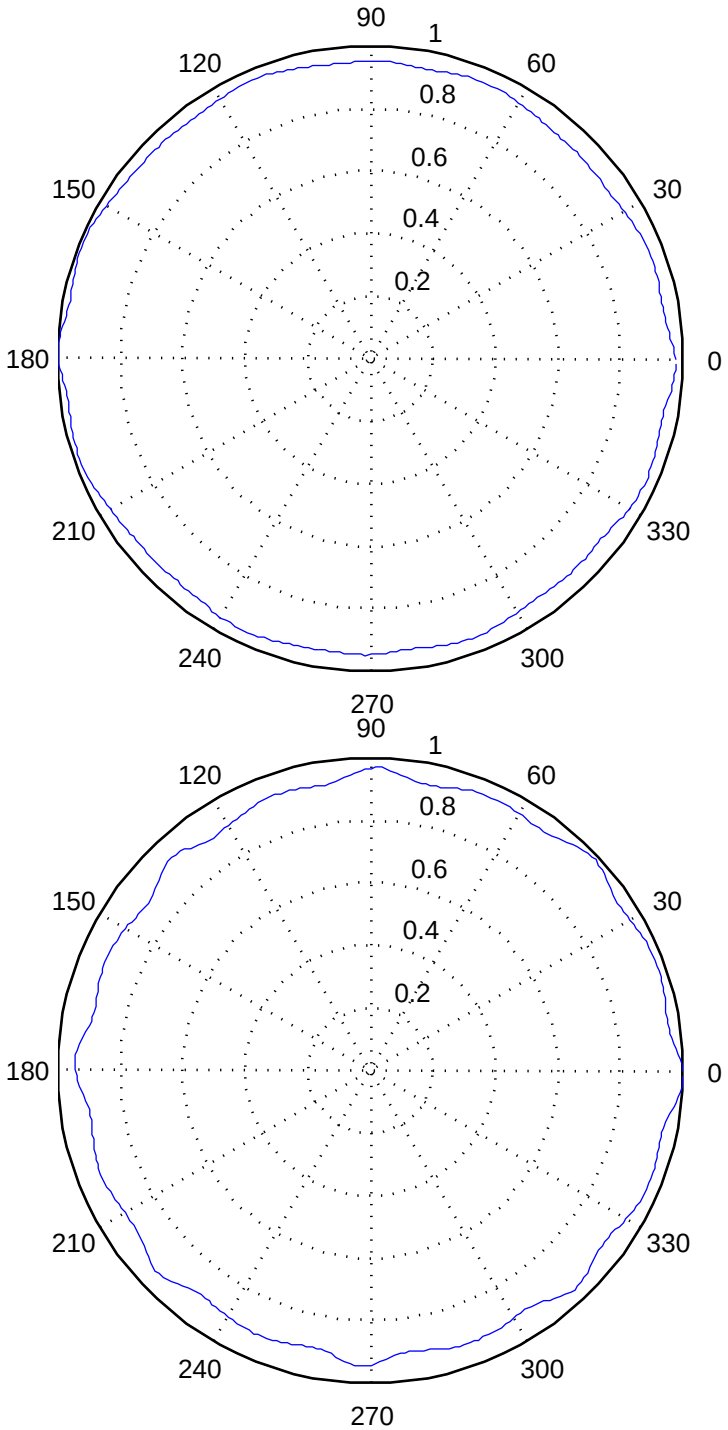

Figure 4: Broadband power-pattern obtained at 2m (top) and 5m (bottom) from center of the room by averaging over all polar angles; the scenario is synthesized in a very large room using 48 loudspeakers.

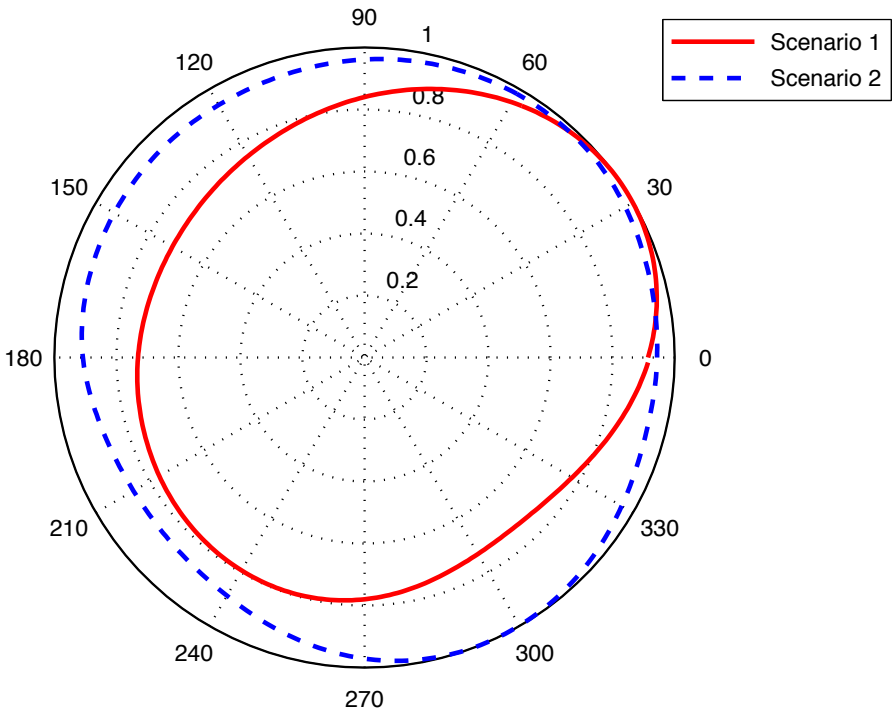

Figure 5: Diffuseness assessment using broadband power-pattern; scenario 1: ambient source diffuse field and scenario 2: boosted power diffuse field by adding additional sources.

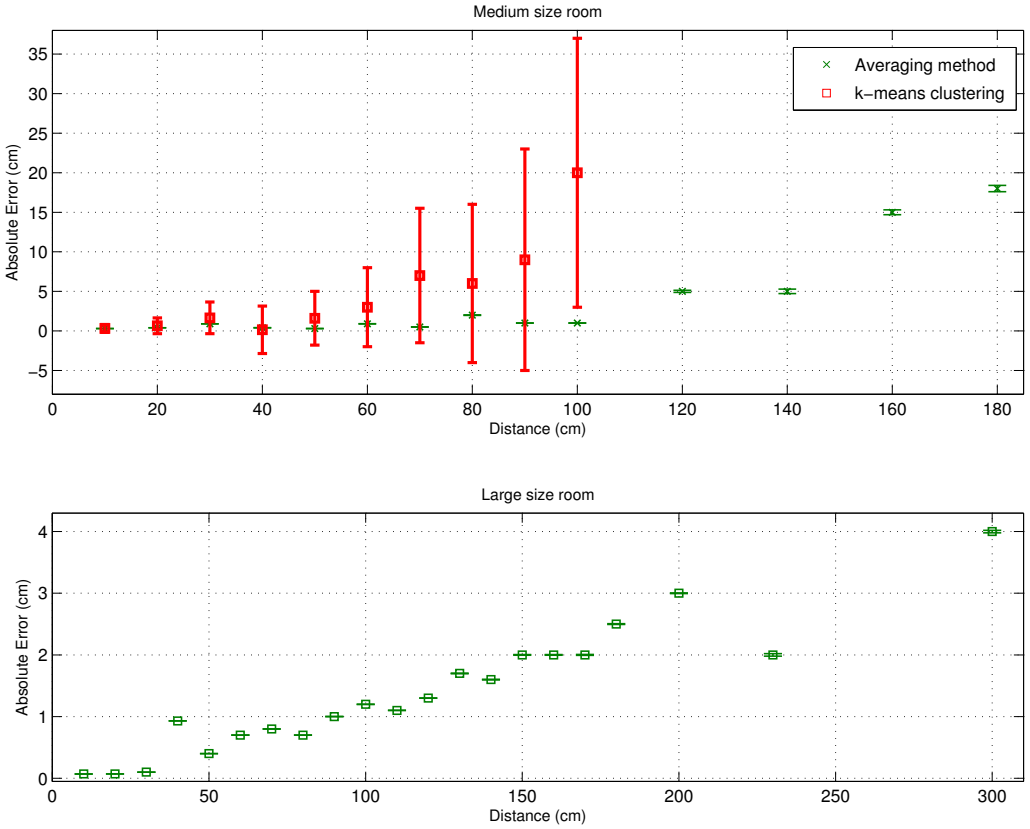

Figure 6: Comparison of error bars for estimation of pairwise distances in the medium (top) and large size (bottom) rooms. In the top plot, "cross" corresponds to the averaging method and "square" corresponds to the k-means clustering. The bottom plot corresponds to the averaging method.

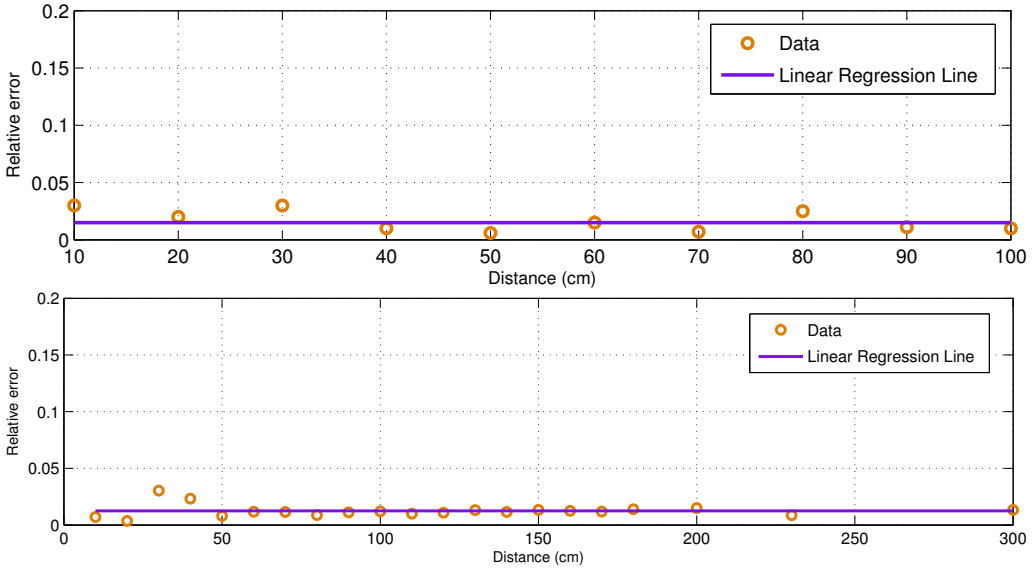

Figure 7: Relative error vs. distance for medium size (top) and large (bottom) rooms. The linear regression can be used to predict the relative error.

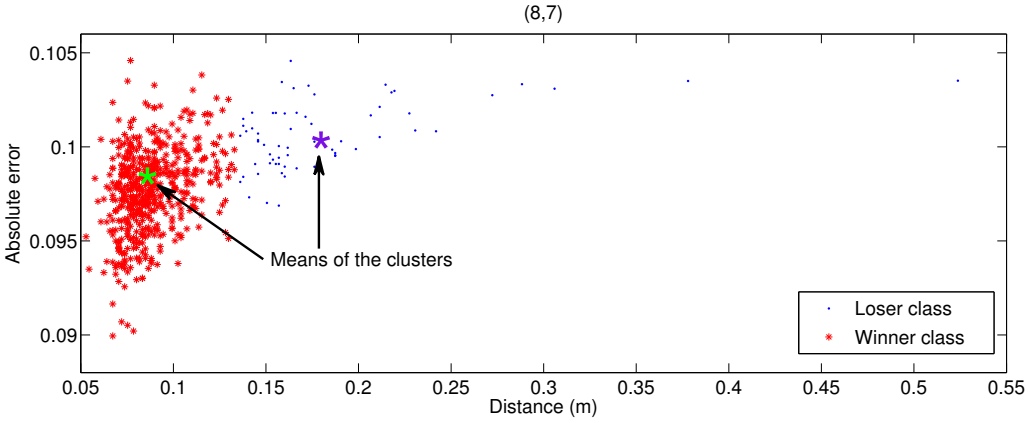

Figure 8: Baseline method: k-means clustering for microphones 7 and 8 located 7.6cm apart. Blue points have high errors and red points are the winners. The estimated pairwise distance is 8.21 cm.

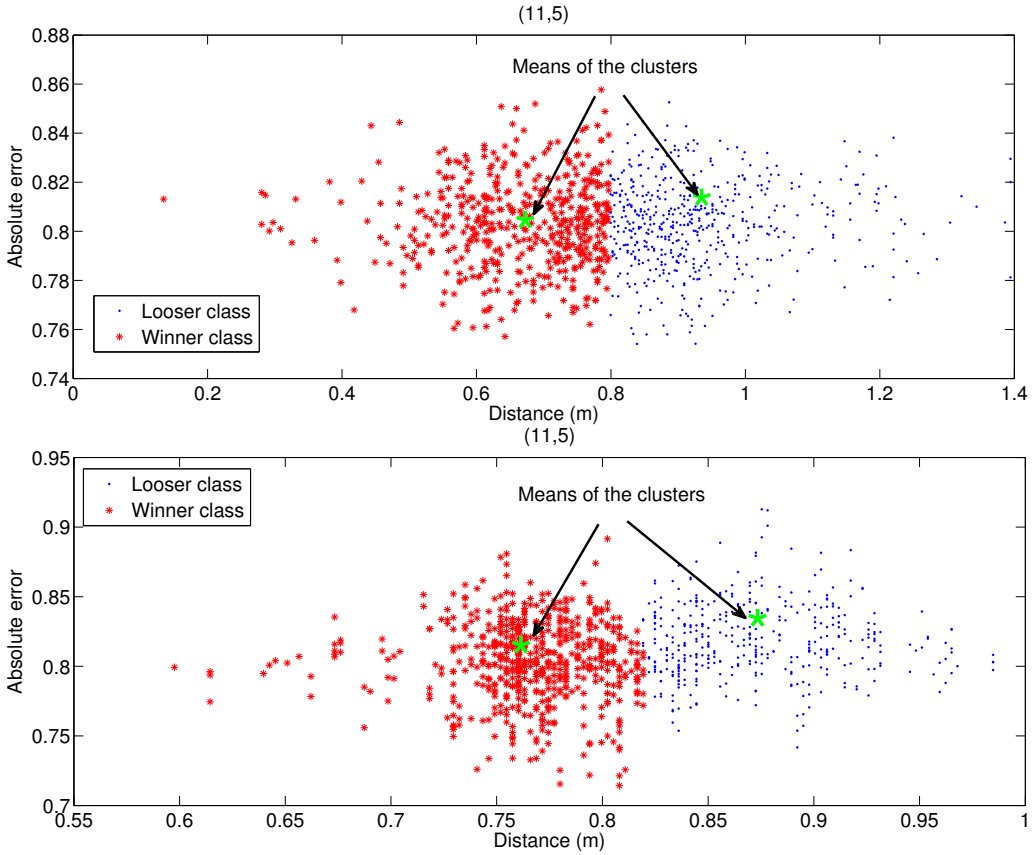

Figure 9: Distance estimation of microphones 11 and 5 using real data recordings. The ground truth is 77.38 cm. (top) Baseline method using k-means clustering on single frame coherence function. The estimated distance is 66 cm. (bottom) k-means clustering on averaged coherence function. The estimated distance is 76.6 cm.

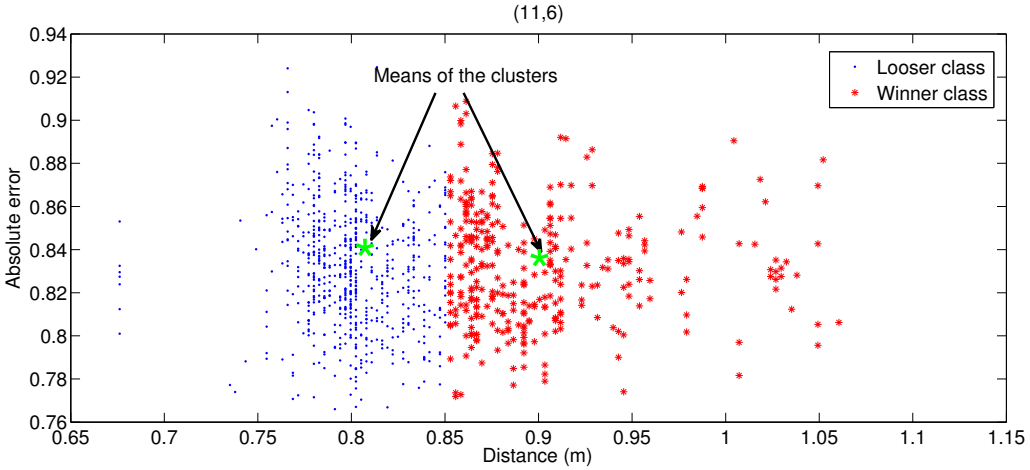

Figure 10: Distance estimation of microphones 11 and 6 using averaging and k-means clustering; correct distance is 80 cm and the estimated distance is 90.2 cm.

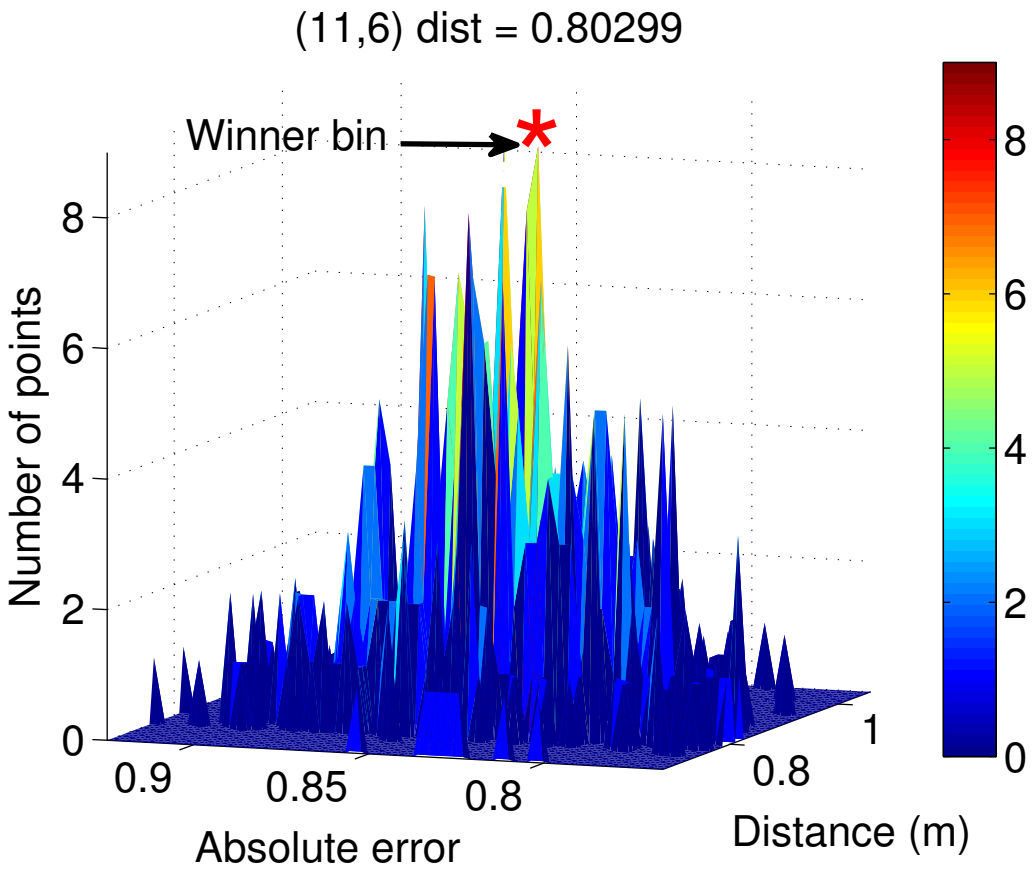

Figure 11: Distance estimation using averaging and two-dimensional histogram clustering; the correct distance is 80 cm and the estimated distance is 80.3 cm.

Distance (cm)	Baseline	BP	AVG+HIS	AVG+BP+HIS	Corresponding microphone pairs as depicted in Figure 3
7.65	.3	.26	.24	.20	$\{(1,2),(2,3),(3,4),(4,5),(5,6),(6,7),(7,8),(8,1)\}$
10	.37	.35	.31	.26	$\{(1,9),(2,9),(3,9),(4,9),(5,9),(6,9),(7,9),(8,9)\}$
14.14	.38	.36	.33	.29	$\{(1,3),(2,4),(3,5),(4,6),(5,7),(6,8),(7,1),(8,2)\}$
18.48	.44	.4	.36	.32	$\{(1,4),(2,5),(3,6),(4,7),(5,8),(6,1),(7,2),(8,3)\}$
20	.55	.47	.45	.35	$\{(1,5),(2,6),(3,7),(4,8)\}$
60	8.4	6.3	3.3	2.7	$\{(4,10),(2,11),(8,12)\}$
70	10.4	9.6	3.5	3.0	$\{(9,10),(9,11),(9,12)\}$
80	14.1	13.5	3.8	3.2	$\{(8,10),(6,11),(4,12)\}$
99	25.2	21.3	4.3	3.6	$\{(10,11),(11,12)\}$

Table 2: Root mean squared error of pairwise distance estimation using diffuse field coherence model evaluated on real data recordings. The presented techniques include the baseline formulation [12], enhanced model by averaging coherence function (AV), using histogram (HIS) for removing the outliers as well as boosting the power (BP) of the sound field.

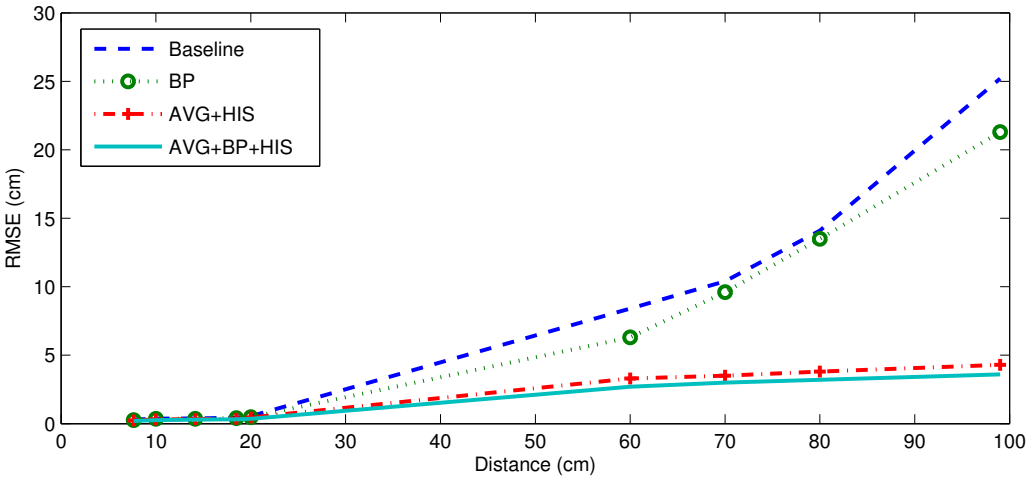

Figure 12: Comparing the performance of all methods using real data recordings for pairwise distance estimation. The baseline is k-means method. BP illustrates the results of using extra broadband sound. Furthermore, AVG+HIS shows big improvement by using averaging method and 2D histogram. Finally AVG+HIS+BP shows the result of applying averaging, histogram and augmenting the sound field.

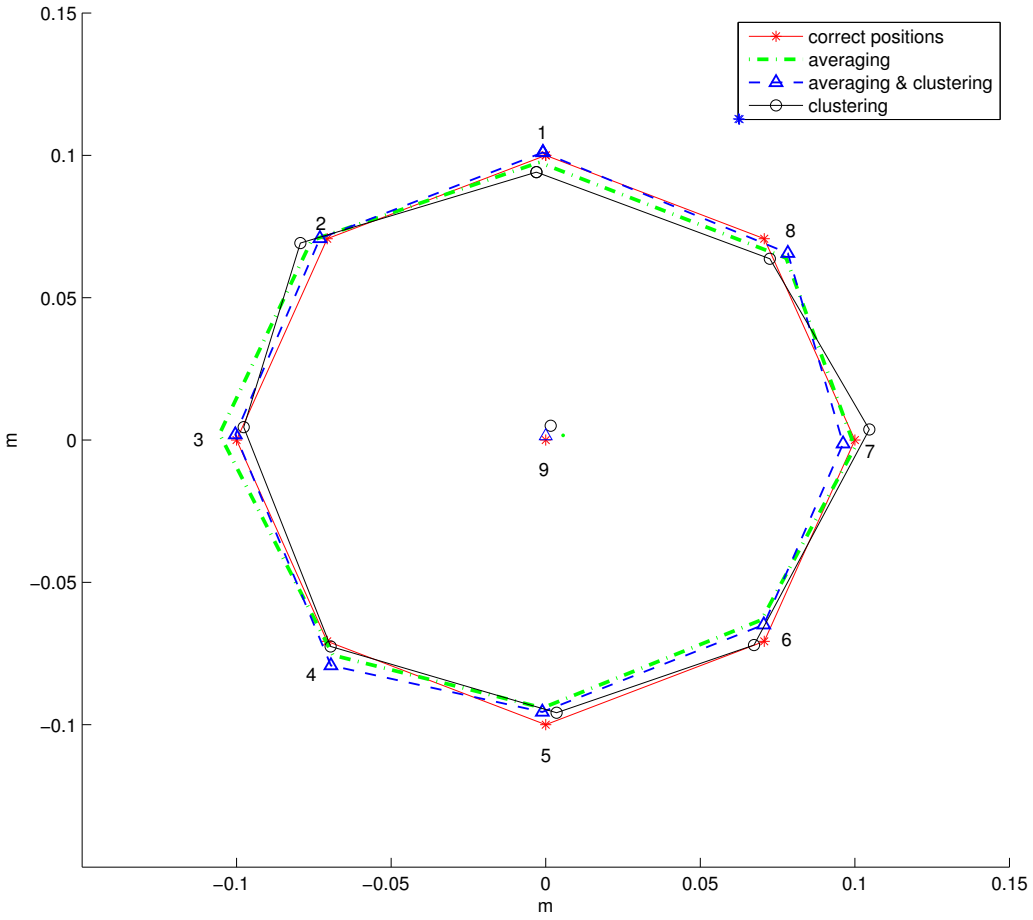

Figure 13: Calibration of a 9-channel microphone array on real diffuse sound field recordings using averaging and a hybrid of averaging and histogram-based clustering.

Method	Error
Baseline	8.83
Averaging	8.04
Averaging + Histogram	5.00

Table 3: Calibration results of 9 microphones.

Room size	ϵ	Max distance (m)
Medium	0.0164	1
Large	0.0124	3
Very large	0.0103	6

Table 4: Maximum pairwise distance that can be estimated with relatively low error in three different rooms based on fitting the sinc function to the average coherence function.

Distance (cm)	RMSE
400	2
500	5
600	10
700	12
800	30
900	25
1000	57

Table 5: Root mean squared error of pairwise distance estimation using diffuse field coherence model for a very large size room (6 times greater than the medium size room).